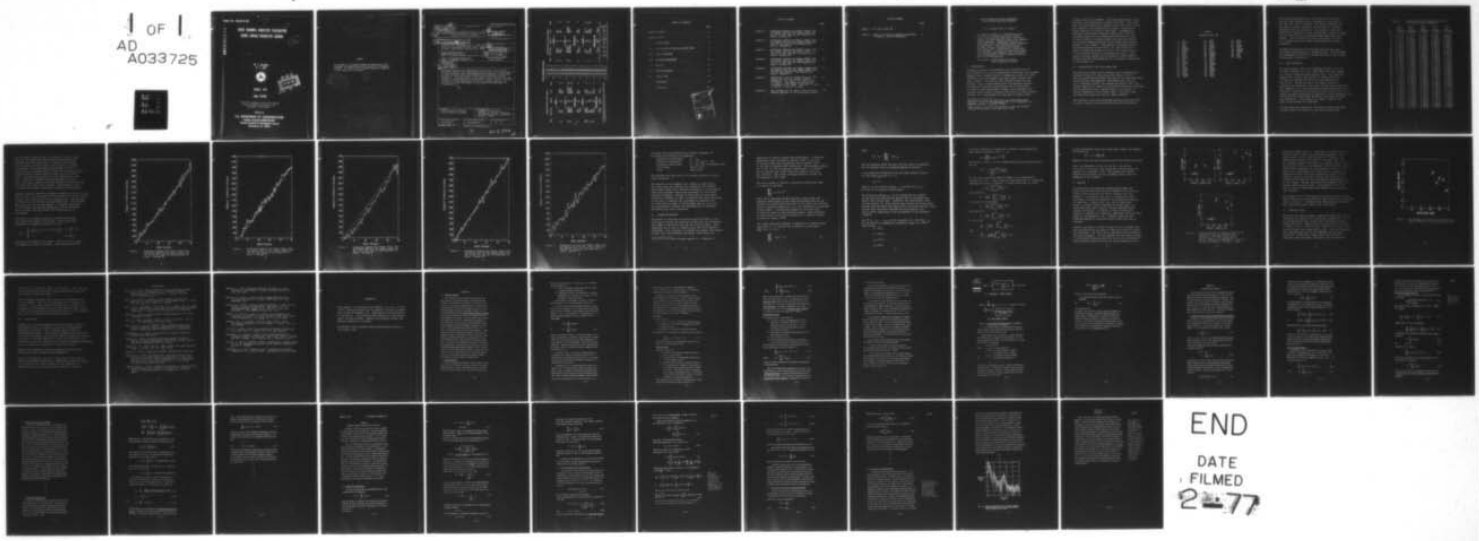
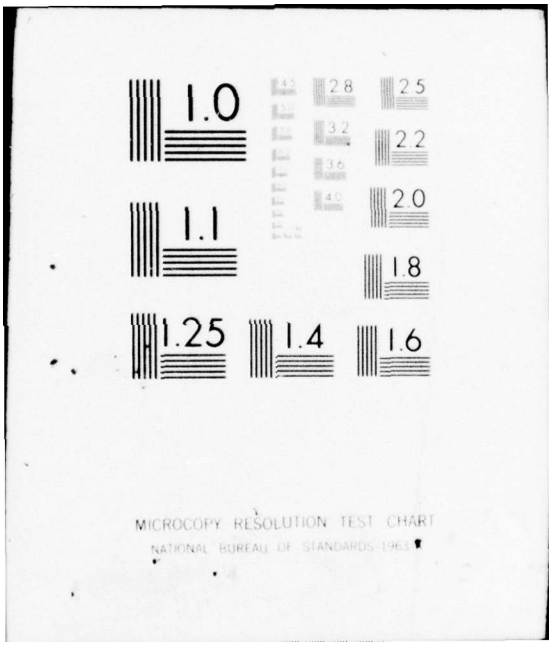
AD-A033 725 INSTITUTE FOR TELECOMMUNICATION SCIENCES BOULDER COLO F/G 17/2
VOICE CHANNEL OBJECTIVE EVALUATION USING LINEAR PREDICTIVE CODI--ETC(U)
AUG 76 W J HARTMAN, S F BOLL DOT-FA74WAI-448

UNCLASSIFIED FAA-RD-75-189 NL

1 OF 1
AD A033725



END
DATE FILMED
2-77


1.0
1.1
1.25
1.4
1.6
2.8
2.5
3.2
2.2
3.6
2.0
4.0
1.8
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963

Report No. FAA-RD-75-189

VOICE CHANNEL OBJECTIVE EVALUATION USING LINEAR PREDICTIVE CODING

W. J. Hartman
S. F. Boll

AUGUST 1976

FINAL REPORT

Prepared for

U.S. DEPARTMENT OF TRANSPORTATION
FEDERAL AVIATION ADMINISTRATION
Systems Research & Development Service
Washington, D.C. 20590

NOTICE

This document is disseminated under the sponsorship of the Department of Transportation in the interest of information exchange. The United States Government assumes no liability for its contents or use thereof.

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
FAA-RD-75-189		

4. Title and Subtitle
Voice Channel Objective Evaluation Using Linear Predictive Coding

5. Report Date
August 1976

6. Performing Organization Code
910.02

7. Author(s)
W.J. Hartman and S.F. Boll

8. Performing Organization Report No.

9. Performing Organization Name and Address
Institute for Telecommunication Sciences,
Office of Telecommunications, Dept. Commerce
Boulder, Laboratories, Boulder, CO 80302

10. Work Unit No. (TRAIS)

11. Contract or Grant No.
DOT-FA74WAI-448

12. Sponsoring Agency Name and Address
Federal Aviation Agency,
Dept. of Transportation, Washington, DC

Office of Telecommunications
Dept. of Commerce, Boulder, CO

13. Type of Report and Period Covered
Final Report.

14. Sponsoring Agency Code
SRDS ARD-60

15. Supplementary Notes
54p.

16. Abstract

We present results of a feasibility study on the use of linear predictive coding (LPC) techniques for deriving an objective measure of intelligibility over voice communication channels. Background material is given and several potentially useful measures are identified. The limitations of the present study are detailed and methods of overcoming these limitations in future work are outlined. In spite of these limitations, the study strongly supports the suitability of LPC techniques for the objective measurement of intelligibility.

17. Key Words	18. Distribution Statement
Intelligibility testing, linear predictive coding.	Document is available through the National Technical Information Service, Springfield, Virginia 22151

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	50	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

A 403394 LB

METRIC CONVERSION FACTORS

Approximate Conversions to Metric Measures

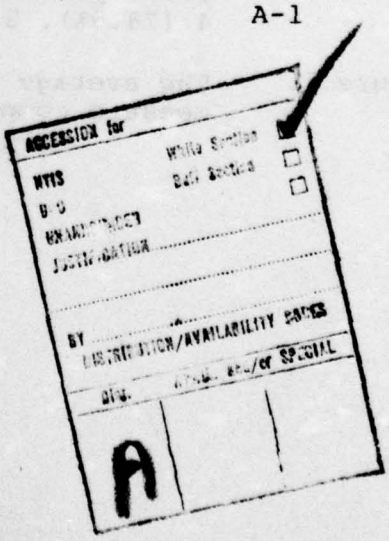
Symbol	When You Know	Multiply by	To Find	Symbol
		LENGTH		
in	inches	*2.5	centimeters	cm
ft	feet	30	centimeters	cm
yd	yards	0.9	meters	m
mi	miles	1.6	kilometers	km
		AREA		
in^2	square inches	6.5	square centimeters	cm^2
ft^2	square feet	0.09	square meters	m^2
yd^2	square yards	0.8	square meters	m^2
mi^2	square miles	2.6	square kilometers	km^2
	acres	0.4	hectares	ha
		MASS (weight)		
oz	ounces	28	grams	g
lb	pounds	0.45	kilograms	kg
	short tons (2000 lb)	0.9	tonnes	t
		VOLUME		
tsp	teaspoons	5	milliliters	ml
Tbsp	tablespoons	15	milliliters	ml
fl oz	fluid ounces	30	milliliters	ml
c	cups	0.24	liters	l
pt	pints	0.47	liters	l
qt	quarts	0.95	liters	l
gal	gallons	3.8	liters	l
ft^3	cubic feet	0.03	cubic meters	m^3
yd^3	cubic yards	0.76	cubic meters	m^3
		TEMPERATURE (exact)		
°F	Fahrenheit temperature	5/9 (after subtracting 32)	Celsius temperature	°C

*1 in = 2.54 (exactly). For other exact conversions and more detailed tables, see NBS Misc. Publ. 286, Units of Weights and Measures, Price $2.25, SD Catalog No. C13.10:286.

Approximate Conversions from Metric Measures

Symbol	When You Know	Multiply by	To Find	Symbol
		LENGTH		
mm	millimeters	0.04	inches	in
cm	centimeters	0.4	inches	in
m	meters	3.3	feet	ft
m	meters	1.1	yards	yd
km	kilometers	0.6	miles	mi
		AREA		
cm^2	square centimeters	0.16	square inches	in^2
m^2	square meters	1.2	square yards	yd^2
km^2	square kilometers	0.4	square miles	mi^2
ha	hectares (10,000 m^2)	2.5	acres	
		MASS (weight)		
g	grams	0.035	ounces	oz
kg	kilograms	2.2	pounds	lb
t	tonnes (1000 kg)	1.1	short tons	
		VOLUME		
ml	milliliters	0.03	fluid ounces	fl oz
l	liters	2.1	pints	pt
l	liters	1.06	quarts	qt
l	liters	0.26	gallons	gal
m^3	cubic meters	35	cubic feet	ft^3
m^3	cubic meters	1.3	cubic yards	yd^3
		TEMPERATURE (exact)		
°C	Celsius temperature	9/5 (then add 32)	Fahrenheit temperature	°F

TABLE OF CONTENTS



LIST OF FIGURES

LIST OF TABLES

VOICE CHANNEL OBJECTIVE EVALUATION USING LINEAR PREDICTIVE CODING

W.J. Hartman* and S.F. Boll**

We present results of a feasibility study on the use of linear predictive coding (LPC) techniques for deriving an objective measure of intelligibility over voice communication channels. Background material is given and several potentially useful measures are identified. The limitations of the present study are detailed and methods of overcoming these limitations in future work are outlined. In spite of these limitations, the study strongly supports the suitability of LPC techniques for the objective measurement of intelligibility.

Key words: Intelligibility testing, linear predictive coding.

I. INTRODUCTION

A common procedure for determining the intelligibility of voice channels is to use a predetermined vocabulary with selected speakers and a listener panel to subjectively grade the intelligibility after the spoken words have passed through some voice channel. A variety of such testing schemes has been devised. Most of these schemes have the desirable property of producing repeatable results which can be interpreted in terms of user requirements. However, the requirement for listener panels greatly restricts the utility of these testing methods, and a long-sought goal has been to replace these

---

*The author is with the Institute for Telecommunication Sciences, Office of Telecommunications, U.S. Department of Commerce, Boulder, Colorado 80302.

**The author is with the University of Utah and Software Sciences Corporation, Salt Lake City, Utah.

listener panels with hardware. The work reported here covers one step in the direction of reaching that goal. This study uses a 50-word phonetically balanced word list played through five different voice systems with a range of articulation scores from 64.7% to 95% as a data base. A mathematical technique called Linear Predictive Coding (LPC) which was originally applied to the analysis and synthesis of voice is used to derive a distance measure for each word between the original undistorted words and the words after passing through a voice channel. This measure is compared with the subjective scoring for each word. For those words with a range of subjective scores (% correct) the distance measure is a decreasing function of the subjective scores. These results strongly support the suitability of using measures derived from the LPC methodology for an objective measurement of intelligibility.

II. DESCRIPTION OF THE VOICE TAPES USED

The word group chosen for the analysis was the phonetically balanced 50-word list (word group 284) shown in table 1. This word group had previously been used by the Army Electronic Proving Ground Electromagnetic Environment Test Facility, at Ft. Huachuca, Arizona, for systems evaluations work, and several tapes with a range of articulation scores were available. The articulation score is defined here as the percent of correct responses by the listener panel. The master tape was made using three male and two female speakers.

The decision to use the prerecorded prescored tapes was made so that a range of articulation scores would be available.

Table 1

PB Word Group 284

1	JELL	21	BIND	41	FOOD
2	SPICE	22	LICK	42	PINT
3	COD	23	CALF	43	ROT
4	CHEW	24	CATCH	44	RHYME
5	THREAD	25	DUMB	45	FLIP
6	SHACK	26	US	46	WHEEZE
7	BOLT	27	FORTH	47	GUESS
8	LOOK	28	YEAST	48	ASK
9	LEFT	29	FROCK	49	FAD
10	DEUCE	30	EACH	50	ROPE
11	BID	31	NIGHT		
12	KILL	32	WIG		
13	CRACK	33	QUEEN		
14	DAY	34	FRONT		
15	TILL	35	ROD		
16	SLIDE	36	EASE		
17	CLOD	37	FREAK		
18	THIS	38	HUM		
19	BORED	39	REST		
20	CHANT	40	ROLL		

The scores chosen were 64.7%, 73%, 78.5%, 87.3%, and 95%. For each tape, information was available for each word on the number of correct responses, the actual responses made, and other pertinent information on the scoring. However, information was not obtained on the specific system which produced the distortion (noise) on the tapes, except to assure that at least some of the distorted tapes came from digital systems. Table 2 lists the word by word scores for each of the tapes, giving the number and percentage of correct responses.

Although the motivation for choosing the tapes was sound, the choice caused problems in aligning words. These problems are discussed in Section III and one solution to this problem is discussed in Section VI.

III. DATA PROCESSING

The analog signals were first processed through an AGC and low pass filtered to 3.2 kHz. This signal was then sampled at a rate of 8 kHz and quantized to 8 bits. The quantized samples were blocked into records 160 samples long and the mean and standard deviation (SD) was calculated for each record. The SD was used as an energy criteria to determine the endpoints of the words, and the word "midpoints". The "midpoint" for a word was defined to be the beginning of the 160-point record which was centered in those records with the largest standard deviations for that word. It was defined in this way so that the correlations described in the next paragraph were meaningful. This midpoint was usually close to the point midway between the endpoints.

For the tapes with distortion, the energy criteria were used to first define the midpoint of the 1st and 50th words.

Table 2. Number of correct responses and percent (in paren.) correct responses.

Word	Tape 2 64.7 Percent 6 listeners	Tape 3 73.0 Percent 8 listeners	Tape 4 78.5 Percent 8 listeners	Tape 5 87.3 Percent 6 listeners	Tape 6 95 Percent 8 listeners
1	6 (100.0)	7 (87.5)	6 (75.0)	6 (100.0)	8 (100.0)
2	0 (0.0)	3 (37.5)	3 (37.5)	4 (66.6)	8 (100.0)
3	6 (100.0)	8 (100.0)	8 (100.0)	5 (83.3)	8 (100.0)
4	5 (83.3)	5 (62.5)	8 (100.0)	6 (100.0)	8 (100.0)
5	3 (50.0)	6 (75.0)	7 (87.5)	6 (100.0)	8 (100.0)
6	3 (50.0)	8 (100.0)	4 (50.0)	5 (83.3)	7 (87.5)
7	2 (33.3)	7 (87.5)	4 (50.0)	6 (100.0)	8 (100.0)
8	6 (100.0)	6 (75.0)	8 (100.0)	6 (100.0)	8 (100.0)
9	1 (16.6)	1 (12.5)	1 (12.5)	5 (83.3)	8 (100.0)
10	6 (100.0)	8 (100.0)	7 (87.5)	6 (100.0)	8 (100.0)
11	2 (33.3)	1 (12.5)	6 (75.0)	5 (83.3)	7 (87.5)
12	6 (100.0)	7 (87.5)	8 (100.0)	5 (83.3)	8 (100.0)
13	6 (100.0)	8 (100.0)	7 (87.5)	6 (100.0)	8 (100.0)
14	4 (66.6)	7 (87.5)	7 (87.5)	4 (66.6)	8 (100.0)
15	5 (83.3)	5 (62.5)	8 (100.0)	4 (66.6)	8 (100.0)
16	3 (50.0)	4 (50.0)	8 (100.0)	6 (100.0)	7 (87.5)
17	6 (100.0)	7 (87.5)	7 (87.5)	4 (66.6)	7 (87.5)
18	3 (50.0)	4 (50.0)	5 (62.5)	3 (50.0)	8 (100.0)
19	4 (66.6)	7 (87.5)	6 (75.0)	5 (83.3)	8 (100.0)
20	1 (16.6)	5 (62.5)	4 (50.0)	5 (83.3)	7 (87.5)
21	2 (33.3)	4 (50.0)	1 (12.5)	2 (33.3)	8 (100.0)
22	1 (16.6)	3 (37.5)	3 (37.5)	3 (50.0)	4 (50.0)
23	5 (83.3)	5 (62.5)	7 (87.5)	5 (83.3)	8 (100.0)
24	1 (16.6)	4 (50.0)	5 (62.5)	5 (83.3)	8 (100.0)
25	1 (16.6)	5 (62.5)	4 (50.0)	6 (100.0)	8 (100.0)
26	0 (0.0)	0 (0.0)	3 (37.5)	4 (66.6)	4 (50.0)
27	5 (83.3)	8 (100.0)	8 (100.0)	6 (100.0)	8 (100.0)
28	4 (66.6)	5 (62.5)	5 (62.5)	6 (100.0)	8 (100.0)
29	5 (83.3)	8 (100.0)	8 (100.0)	6 (100.0)	8 (100.0)
30	5 (83.3)	8 (100.0)	7 (87.5)	6 (100.0)	8 (100.0)
31	6 (100.0)	8 (100.0)	8 (100.0)	5 (83.3)	8 (100.0)
32	0 (0.0)	7 (87.5)	7 (87.5)	5 (83.3)	8 (100.0)
33	6 (100.0)	8 (100.0)	8 (100.0)	6 (100.0)	8 (100.0)
34	3 (50.0)	6 (75.0)	6 (75.0)	6 (100.0)	8 (100.0)
35	6 (100.0)	7 (87.5)	8 (100.0)	6 (100.0)	8 (100.0)
36	2 (33.3)	6 (75.0)	4 (50.0)	3 (50.0)	8 (100.0)
37	5 (83.3)	4 (50.0)	5 (62.5)	5 (83.3)	6 (75.0)
38	1 (16.6)	2 (25.0)	7 (87.5)	4 (66.6)	6 (75.0)
39	4 (66.6)	6 (75.0)	7 (87.5)	6 (100.0)	7 (87.5)
40	6 (100.0)	8 (100.0)	8 (100.0)	6 (100.0)	8 (100.0)
41	5 (83.3)	6 (75.0)	7 (87.5)	6 (100.0)	8 (100.0)
42	6 (100.0)	8 (100.0)	8 (100.0)	6 (100.0)	8 (100.0)
43	4 (66.6)	3 (37.5)	3 (37.5)	6 (100.0)	7 (87.5)
44	6 (100.0)	8 (100.0)	8 (100.0)	5 (83.3)	8 (100.0)
45	1 (16.6)	5 (62.5)	8 (100.0)	6 (100.0)	8 (100.0)
46	6 (100.0)	8 (100.0)	8 (100.0)	6 (100.0)	8 (100.0)
47	5 (83.3)	8 (100.0)	7 (87.5)	6 (100.0)	8 (100.0)
48	6 (100.0)	8 (100.0)	8 (100.0)	6 (100.0)	8 (100.0)
49	6 (100.0)	6 (75.0)	8 (100.0)	6 (100.0)	8 (100.0)
50	3 (50.0)	6 (75.0)	8 (100.0)	6 (100.0)	7 (87.5)

Various length intervals about the midpoint of word 1 were then used to obtain the cross correlation against similar intervals from the master word 1, adjusting the midpoint for the distorted word to agree with the point of maximum correlation. Using the midpoints from words 1 and 50, a linear function was determined to relate the distances between midpoints from the master words to those for the distorted words. The correlation process was then repeated for each word. Although no problems were encountered in aligning many of the words, some of the words presented special difficulties which were never fully resolved. The procedures given in Section VI describe a method for obtaining synchronization.

Figures 1 through 5 show the difference between the midpoint of the distorted word and the midpoint of the master word for each of five systems. The linear trend is to be expected due to normal speed differences of the 1/4-inch voice recorders used. The variation about this line was not anticipated. A possible explanation is that these tapes had been rerecorded, and the cumulative tape stretching, wow and flutter all contributed to the differences.

Following the alignment procedure, the data were blocked into frames consisting of 256 points. These frames were windowed with a Hamming window of the form

$$W_H(t) = \begin{cases} \dfrac{1}{1.08\ T_W}\left(0.54 + 0.46 \cos \dfrac{\pi t}{T_W}\right) & \text{for } |t| \leq T_W \\ 0 & \text{elsewhere} \end{cases}$$

where T_W is the width of the window. This was used to insure the stability of the LPC analysis. These windows were then

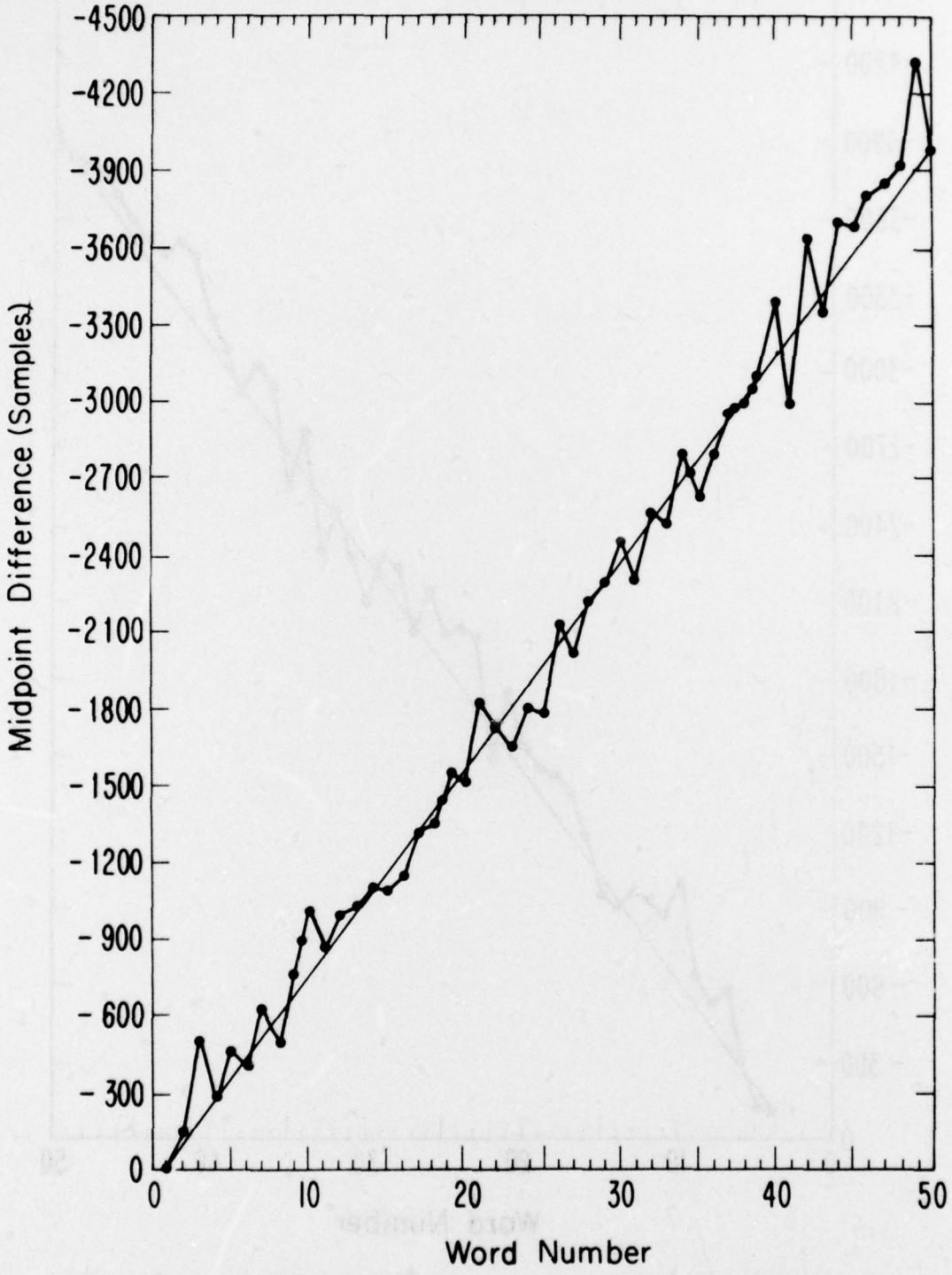


Figure 1. Difference between the sample number from the master tape and the sample number for the 64.7% (#2) tape for the midpoints of words 1 through 50.

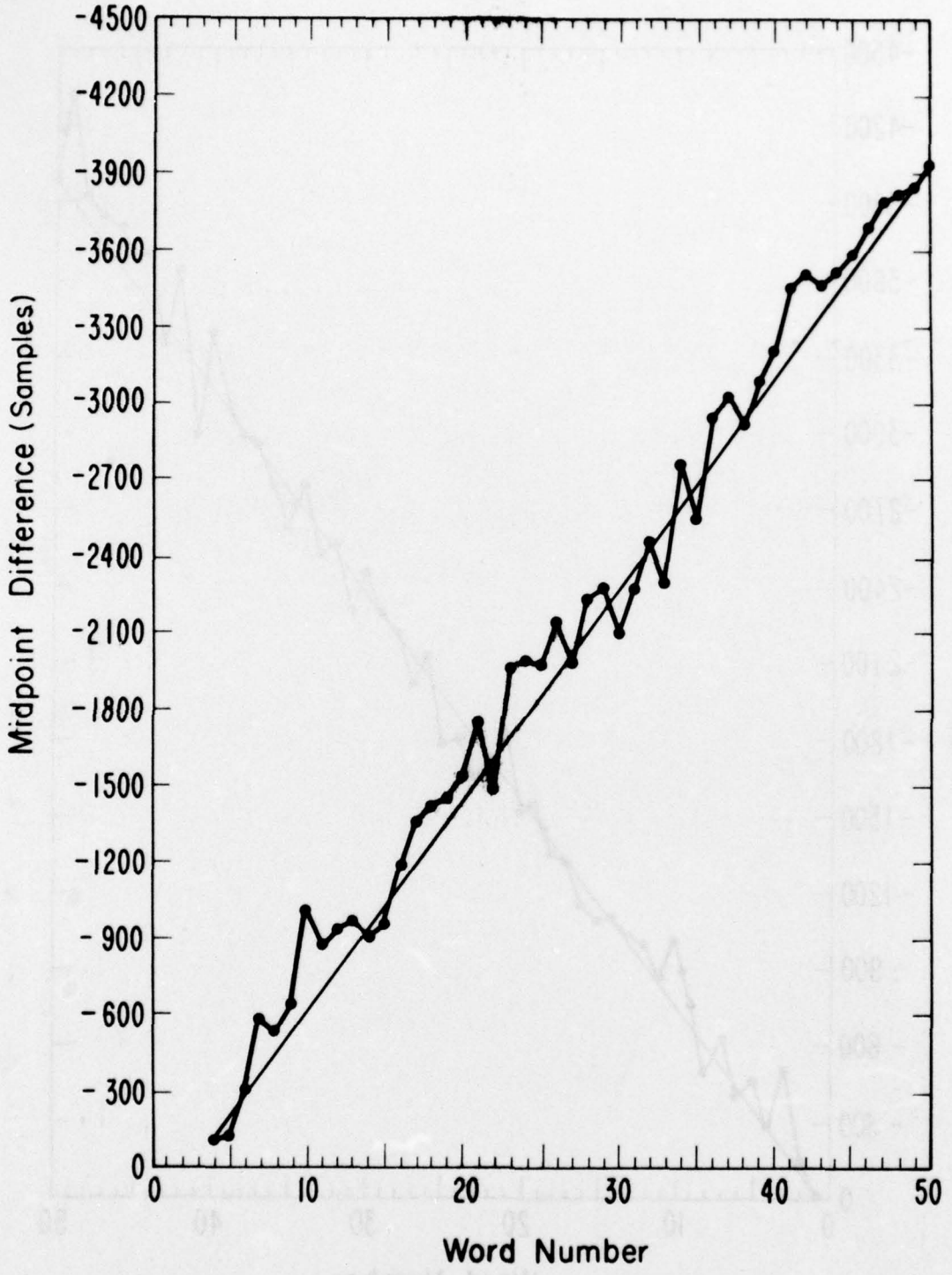


Figure 2. Difference between the sample number from the master tape and the sample number for the 73% (#3) tape for the midpoints of words 1 through 50.

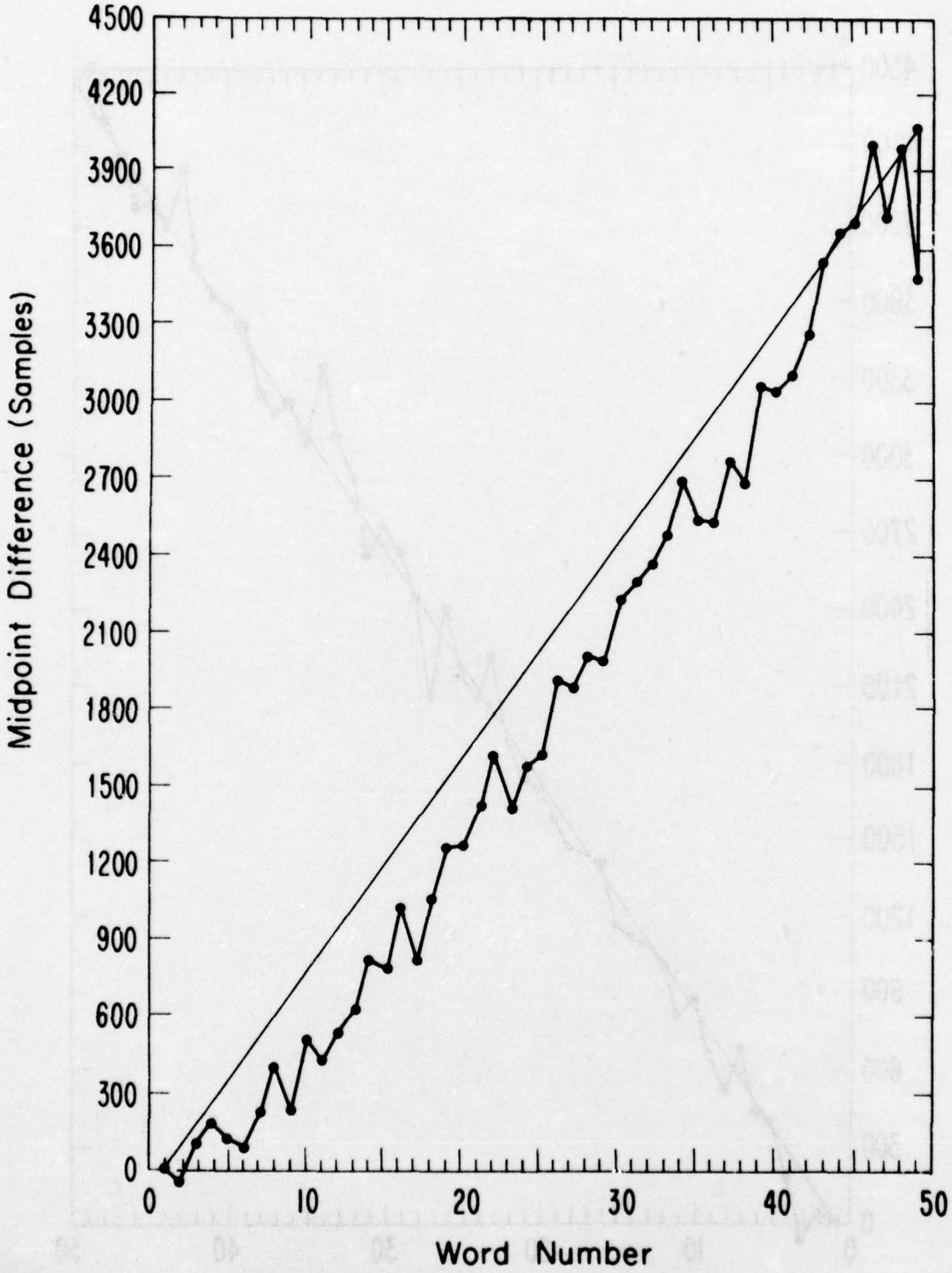


Figure 3. Difference between the sample number from the master tape and the sample number for the 78.5% (#4) tape for the midpoints of words 1 through 50.

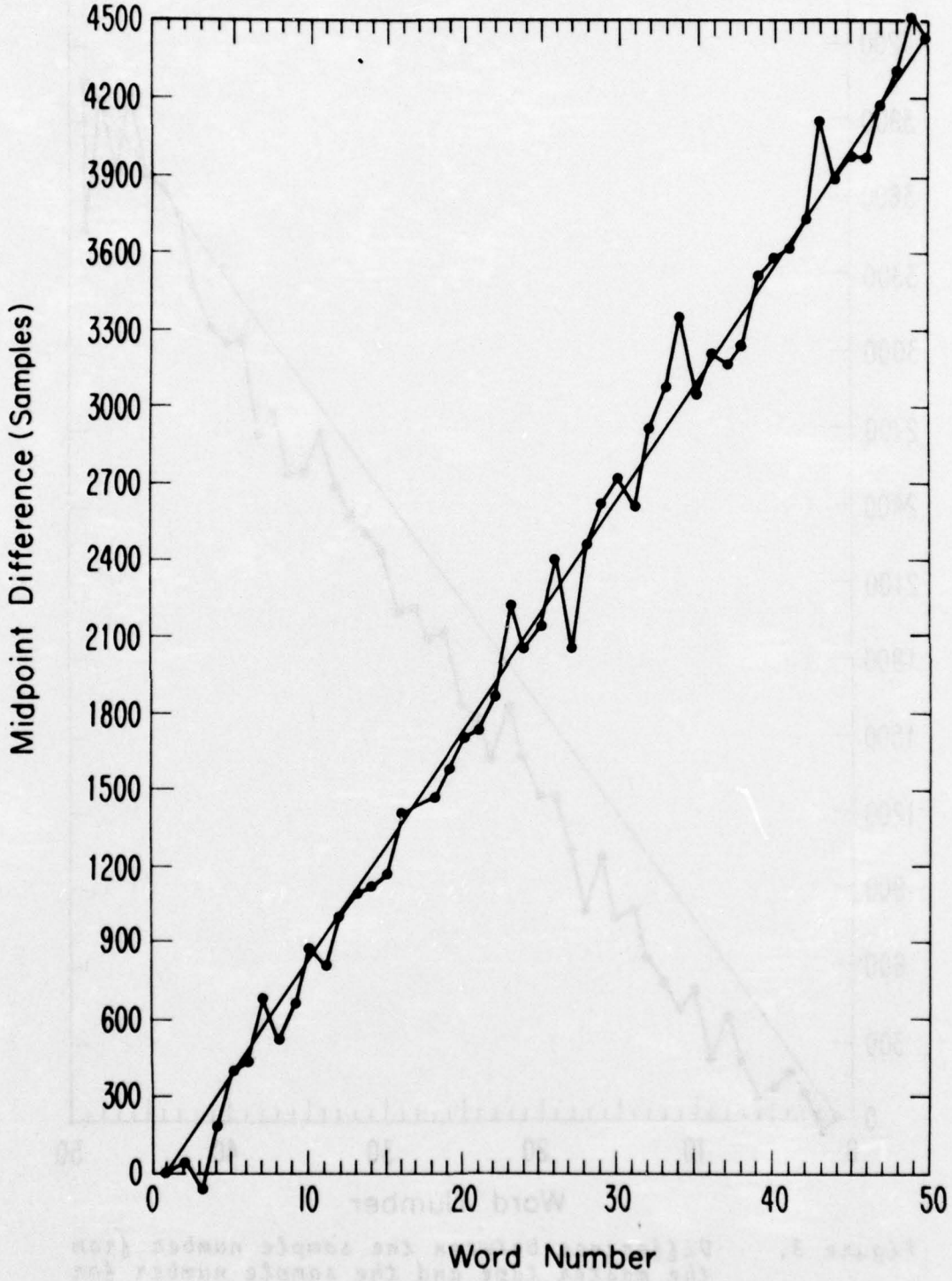


Figure 4. Difference between the sample number from the master tape and the sample number for the 87.3% (#5) tape for the midpoints of words 1 through 50.

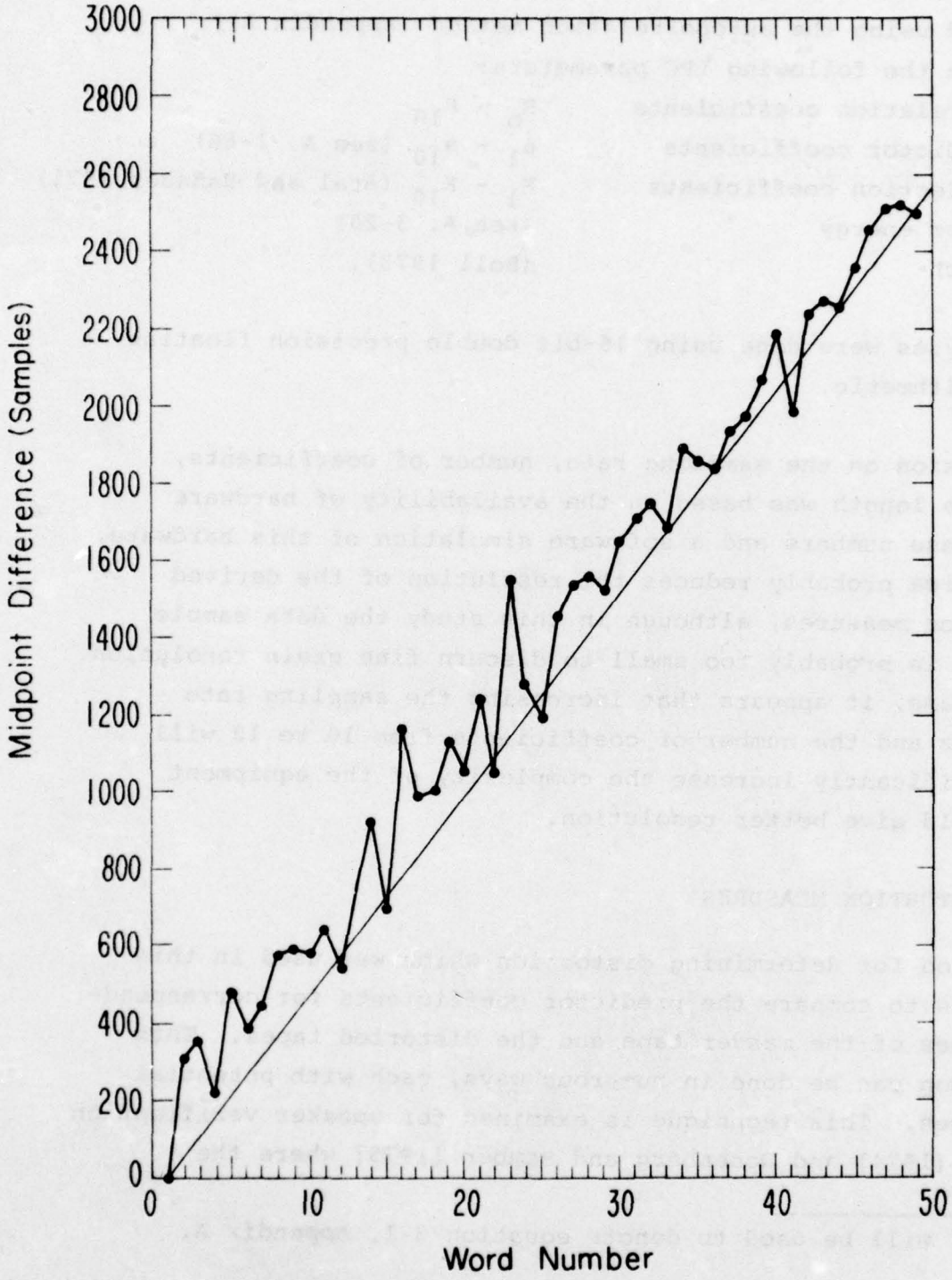


Figure 5. Differences between the sample number from the master tape and the sample number for the 95% (#6) tape for the midpoints of words 1 through 50.

processed using the autocorrelation method (Appendix A)* to obtain the following LPC parameters:

Correlation coefficients	R_o - R_{10}
Predictor coefficients	a_1 - a_{10} (see A, 1-66)
Reflection coefficients	K_1 - K_{10} (Atal and Hanauer 1971)
Error energy	(see A, 3-20)
Pitch	(Boll 1973).

The analyses were done using 16-bit double precision floating point arithmetic.

The decision on the sampling rate, number of coefficients, and frame length was based on the availability of hardware using these numbers and a software simulation of this hardware. This choice probably reduces the resolution of the derived distortion measures, although in this study the data sample analyzed is probably too small to discern fine grain resolution. In any case, it appears that increasing the sampling rate to 10 kHz and the number of coefficients from 10 to 12 will not significantly increase the complexity of the equipment and should give better resolution.

IV. DISTORTION MEASURES

The method for determining distortion which was used in this study was to compare the predictor coefficients for corresponding frames of the master tape and the distorted tapes. This comparison can be done in numerous ways, each with potential advantages. This technique is examined for speaker verification by Atal [1974] and Rosenberg and Sambur [1975] where the

*(A,3-1) will be used to denote equation 3-1, Appendix A.

comparison is directly between the coefficients. A different approach has been developed by Makhoul [1973] in the area of variable frame rate transmission and by Itakura [1975] in the area of isolated word recognition. These methods examine how "close" one set of predictor coefficients is to another set by comparing the linear prediction residual resulting from each set. This latter technique proved to be best for the purposes of this study.

The first attempt to identify a distortion measure was based on a metric of the form

$$\sum_{i=1}^{p} w_i \, (a_i - a_i')^2$$

where the w_i represent weights and the a_i were either the predictor coefficients or the reflection coefficients and the unprimed and primed quantities refer to the master and distorted signal parameters respectively as it will throughout the remainder of this report. No distortion measure of this form could be found to relate to the subjective scores. Some discussion of the failure of this form of measure in a different context is given by Itakura [1975].

The linear prediction residual is defined for a sampled signal $\{s_n\}$ (n=0,1,2...N) and any set of predictor coefficients $\{a_k'\}$ (k=1, 2...p) as

$$\sum_{n=0}^{N} (e_n')^2 = D_1 \qquad (1)$$

where

$$e'_n = s_n - \sum_{k=1}^{p} a'_k s_{n-k} \quad .$$

For this analysis N=256 and p=10, and the limits of summation will be omitted since no misunderstanding is involved.

If the predictor coefficients are the least squares solution for the signal $\{s_n\}$ then (A, 3-3)

$$D_1 = E,$$

where E is the minimum residual. It follows that $D_1 \geq E$ for any set of predictor coefficients.

The distortion measure D_1, can be interpreted as a measure of how close the coefficients derived from the distorted data predict the original data. Similarly, one could define a distortion measure, D_2, which measures how closely the coefficients derived from the master data $\{a_k\}$ predict the distorted data, $\{s'_n\}$. Several other distortion measures are also possible, and can be easily summarized by the introduction of the following notation.

Let $a^T=(1, -a_1, \ldots -a_{10})$ be the transpose of a, and $R=\{R_{|i-j|}\}$ be the matrix of unnormalized correlation values (A, 3-16). Then we have

$$E = a^T R\, a$$

$$E' = a'^T R' a'$$

$$D_1 = a'^T R\, a'$$

$$D_2 = a^T R' a.$$

It is not difficult to relate each of these to the appropriate power spectra estimates (A,4-7). Let

$$P(\omega)=\left|\sum_{n} W(nT)e^{-jn\omega T}\right|^2$$

be the power spectrum of the windowed data where $W(nT)=W_H(nT)S(nT)$ and let

$$\hat{P}(\omega) = \frac{A^2}{\left|1-\sum_{k=1}^{P} a_k \quad e^{-jk\omega T}\right|^2}$$

be the linear prediction spectrum where A is an appropriate constant (A, 4-6a). Then, using the same convention for primed and unprimed quantities, as before, it can be shown that (A, 4-16)

$$E = \frac{A^2T}{2\,\Pi}\int_{-\Pi/T}^{\Pi/T}\frac{P(\omega)}{\hat{P}(\omega)}\,d\omega\ ,$$

and similarly,

$$E' = \frac{A'^2T}{2\,\Pi}\int_{-\Pi/T}^{\Pi/T}\frac{P'(\omega)}{\hat{P}'(\omega)}\,d\omega\ .$$

It can also be shown that

$$D_1 = \frac{A'^2T}{2\Pi}\int_{-\Pi/T}^{\Pi/T}\frac{P(\omega)}{\hat{P}'(\omega)}\,d\omega\ ,$$

and similarly,

$$D_2 = \frac{A^2T}{2\,\Pi}\int_{-\Pi/T}^{\Pi/T}\frac{P'(\omega)}{\hat{P}(\omega)}\,d\omega.$$

We also have

$$\frac{D_1}{E} = \frac{T\,A'^2}{2\Pi\,A^2}\int_{-\Pi/T}^{\Pi/T}\frac{\hat{P}(\omega)}{\hat{P}'(\omega)}\,d\omega$$

and

$$\frac{D_2}{E'} = \frac{T\,A^2}{2\Pi\,A'^2}\int_{-\Pi/T}^{\Pi/T}\frac{\hat{P}'(\omega)}{\hat{P}(\omega)}\,d\omega\ .$$

Of all the measures which were extensively tested, the measure ASI defined by

$$ASI = 10 \log \left[D_1/E \right]$$

appears to give the best correlation with the subjective scores.

After the completion of most of the work, the measure $5 \log D_1/E + 5 \log D_2/E'$ was tested for a limited sample and appears to be superior to ASI. This measure and related measures are discussed in Gray and Markel [1976] where the asymmetry in ASI is shown.

V. RESULTS

For each word, the distortion measures (ASI) between the master and each distorted tape were computed for 24 analysis frames, twelve preceding and twelve following the word midpoint. An average of ASI over the twenty-four frames was formed, and this was compared to the subjective scoring for that word. Figures 6(a), (b), and (c) show plots of ASI vs percent correct responses for three words. For those words which had a range of correct responses these figures are typical. Of course, a range of distortion measures is possible for completely understood words, and this range is word dependent (i.e., different words can have different amounts of distortion before intelligibility is affected).

Ideally, the points in figure 6 would lie on a straight line, and a detailed study was made on these three words to explain the deviations. The first factor considered was the midpoint and frame alignment. ASI was computed for the analysis frames of the distorted words shifted from the original alignment by $\pm$128 points and $\pm$256 points and the average compared to the

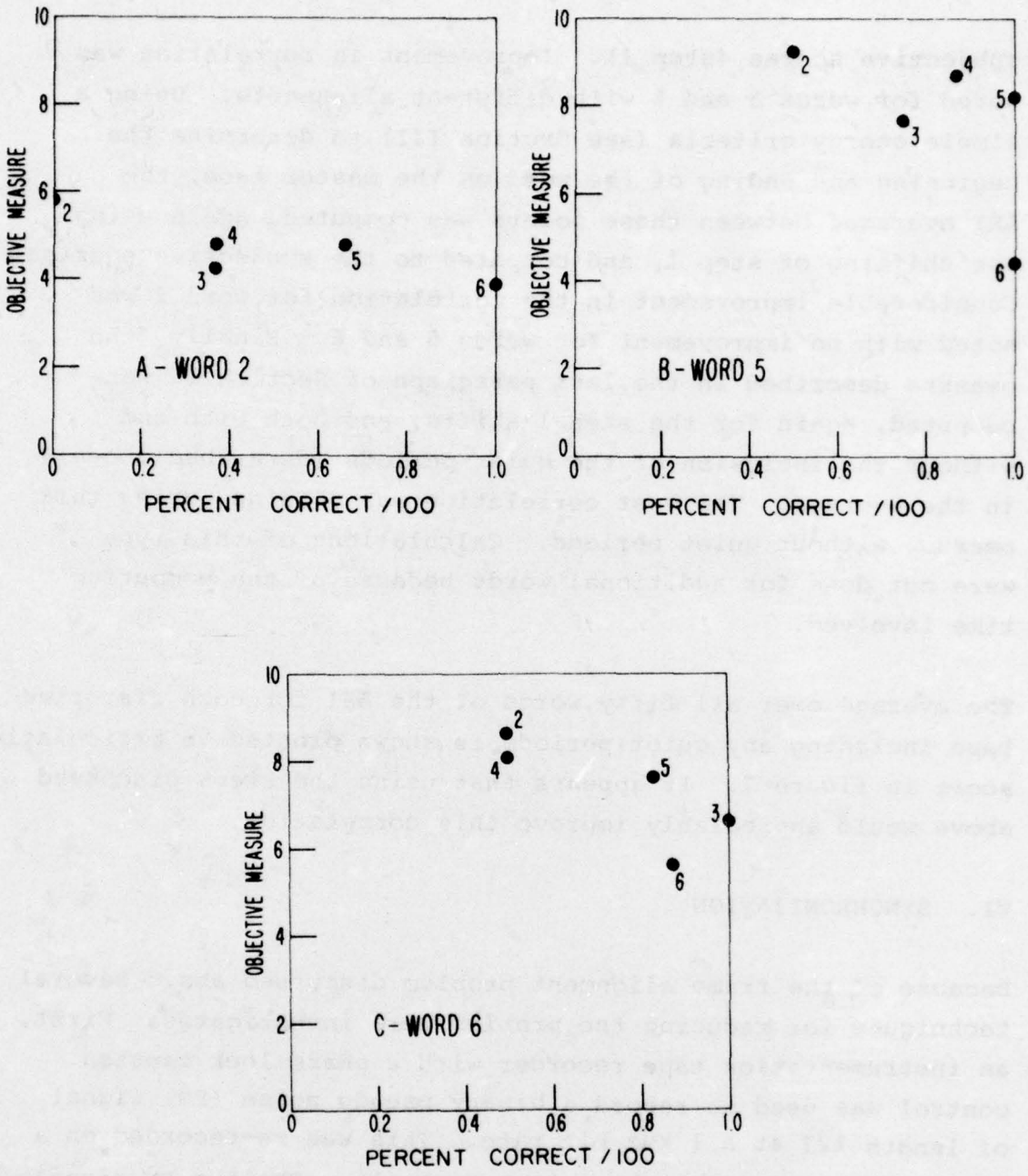


Figure 6. *Comparison of the 24 frame average of the measure ASI with the percent correct responses: (a) word 2, (b) word 5, and (c) word 6. The numbers by the points correspond to the tapes: 2 (64.7%), 3 (73%), 4 (78.5%), 5 (87.3%), 6 (95%).*

subjective scores (step 1). Improvement in correlation was noted for words 5 and 6 with different alignments. Using a simple energy criteria (see Section III) to determine the beginning and ending of the word on the master tape, the ASI averaged between those points was computed, again using the shifting of step 1, and compared to the subjective scoring. Considerable improvement in the correlation for word 2 was noted with no improvement for words 5 and 6. Finally, the measure described in the last paragraph of Section IV was computed, again for the step 1 shifts, and both with and without the inclusion of the quiet periods (described above) in the average. The best correlation was obtained using this measure without quiet periods. Calculations of this type were not done for additional words because of the computing time involved.

The average over all fifty words of the ASI for each distorted tape including any quiet periods is shown plotted vs articulation score in figure 7. It appears that using the steps discussed above would appreciably improve this correlation.

VI. SYNCHRONIZATION

Because of the frame alignment problem discussed above several techniques for reducing the problem were investigated. First, an instrumentation tape recorder with a phase lock capstan control was used to record a binary pseudo noise (PN) signal of length 127 at a 1 kHz bit rate. This was re-recorded on a second recorder (with phase lock control). The two PN signals were then digitized at a 10 kHz sampling rate and cross correlated. Aside from a linear shift due to slightly differing local oscillator frequencies, the two sequences

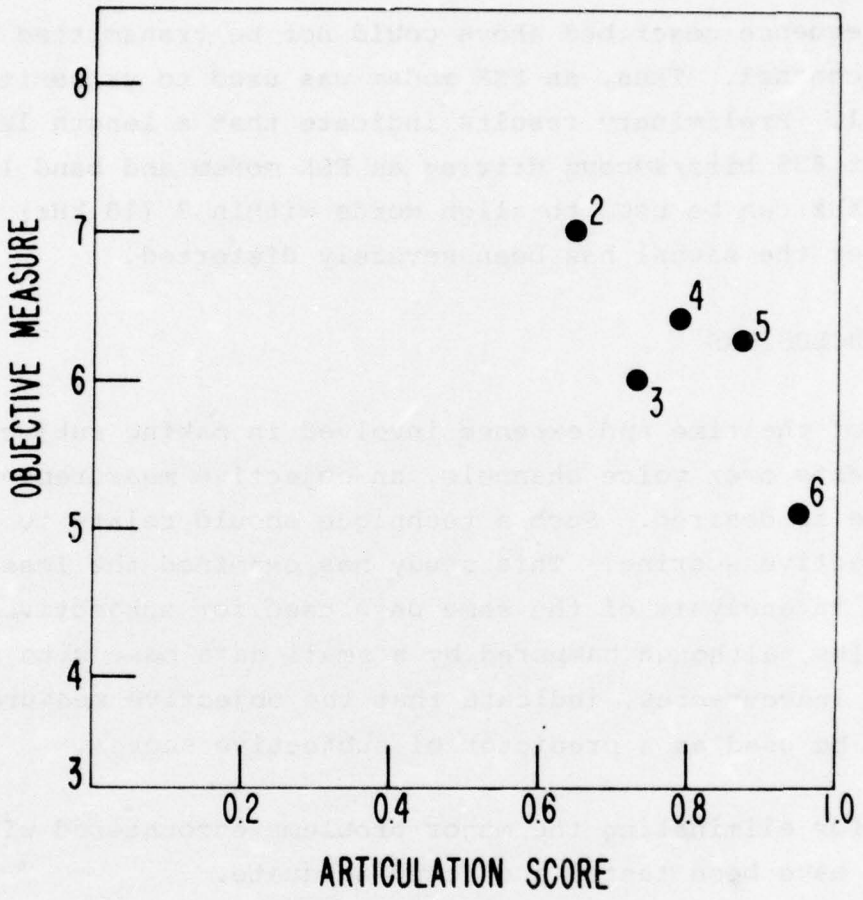


Figure 7. The average over 50 words of the objective measure compared to the articulation score.

differed by less than one sample in 10^7 samples. This indicates that the digitized words from a master and distorted tape can be aligned to within one sample.

The PN sequence described above could not be transmitted over a voice channel. Thus, an FSK modem was used to transmit the PN signal. Preliminary results indicate that a length 127 PN signal at 635 bits/second driving an FSK modem and band limited to 2500 kHz can be used to align words within 3 (10 kHz) samples, even after the signal has been severely distorted.

VII. CONCLUSIONS

Because of the time and expense involved in making subjective measurements over voice channels, an objective measurement technique is desired. Such a technique should relate to the subjective scoring. This study has examined the feasibility of using an analysis of the same data used for subjective scoring. The results, although hampered by a small data base with some inherent inaccuracies, indicate that the objective measure developed here can be used as a predictor of subjective scores.

Methods for eliminating the major problems encountered with the data have been tested and found adequate.

Several refinements of the basic measure were tested and were found to give better predictions. Additional refinements appear to be possible for a small increase in the complexity of computations once the accurate frame alignment is achieved.

REFERENCES

Atal, B.S. (1974), Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification, J. Acoust. Soc. Am., 55, 6, 1304-1312, June.

Atal, B.S., and S.L. Hanauer (1971), Speech analysis and synthesis by linear prediction of the speech wave, J. Acoust. Soc. Am., 50, No. 2 (Part 2), 637-655.

Bell, C.G., H. Fujisaki, J.M. Heinz, K.N. Stevens, and A.S. House (1961), Reduction of speech spectra by analysis-by-synthesis techniques, J. Acoust. Soc. Am., 33, No. 12, 1725-1736.

Boll, S.F. (1973), A prior digital speech analysis, PhD Thesis, University of Utah.

Fant, G. (1960), Acoustic theory of speech production, (Mouton & Co., 's-Gravenhage, The Netherlands).

Gray, A.H.,Jr. and J.D. Markel (1976), Distance measures for speech processing, Private communications, (IEEE Trans. Acoust. Speech and Signal Processing), to be published.

Hildebrand, F.B. (1956), Introduction to numerical analysis, (McGraw-Hill, New York).

Itakura, F. (1975), Minimum prediction residual principle applied to speech recognition, IEEE Trans. Acoust. Speech, and Signal Processing, ASSP-23, 67-72.

Koenig, W., H.K. Dunn, and L.Y. Lacey (1946), The sound spectrograph, J. Acoust. Soc. Am., 18, 19-49.

Kunz, K.S. (1957), Numerical Analysis, (McGraw-Hill, New York).

Levinson, N. (1947), The Wiener RMS (root mean square) error criterion in filter design and prediction, J. Math Phys., 25, No. 4, 261-278. Also in Appendix B of Extrapolation and Smoothing of Stationary Time Series, by N. Wiener, MIT Press, Cambridge, MA, 1966.

Lummis, Robert C. (1973), Speaker verification by computer using speech intensity for temporal registration, IEEE Trans. Audio and Electroacoustics, Vol. AU-21, No. 2, 80-89, April.

Makhoul, J. (1973), Spectral analysis of speech by linear prediction, IEEE Trans. on Audio and Electroacoustics, 140-148, June.

Makhoul, J.I. and J.J. Wolf (1972), Linear prediction and the spectral analysis of speech, BBN Report No. 2304, Cambridge, MA.

Markel, J.D. (1972), Digital inverse filtering - a new tool for formant trajectory estimateion, IEEE Trans. Audio & Electroacoustics, AU-20, No. 2, 129-137. Also, SCRL Monograph 7, SCRL, Santa Barbara, CA, 1971).

Markel, J.D. and A.H. Gray, Jr. (1973), On autocorrelation equations as applied to speech analysis, IEEE Trans. Audio & Electroacoustics, Vol. AU-21, No. 2, 69-79, April.

Mathews, M.V., J.E. Miller, and E.E. David (1961), Pitch synchronous analysis of voiced sounds, J. Acoust. Soc. Am., 33, 179-186.

Noll, A.M. (1964), Short-time spectrum and cepstrum techniques for vocal tract, J. Acoust. Soc. Am., 36, 296-302.

Rosenberg, A.E. and M.R. Sambur (1975), New techniques for automatic speaker verification, IEEE Trans. Acoust., Speech, and Sig. Process., Vol. ASSP-23, No. 2, 169-176, April.

Treitel, S. and E.A. Robinson (1967), Introduction, special issue on the MIT geophysical analysis group reports, Geophysics, 32, No. 3, 416-417.

Wilkinson, J.H. and C. Reinsch (1971), Handbook for automatic computation, 2, Linear Algebra, (Springer-Verlag, New York).

APPENDIX A

This appendix is reproduced from Makhoul, J.L., and J.J. Wolf (1972), Linear prediction and the spectral analysis of speech, Bolt Beranek and Newman, Inc., Cambridge, MA with permission of the authors. Dots indicate where material of the origianl report has been omitted. However, the original chapter and equation numbering is retained.

The marginal notes indicate points particularly pertinent to the present study.

INTRODUCTION

1.1 Historical Overview

One of the most important methods of speech analysis has been the use of the short-time spectrum. This has been accomplished in different ways and to different ends during the past 25 years. The first major breakthrough was the invention of the sound spectrograph (Koenig, Dunn and Lacey, 1946) which is still used extensively for the spectral analysis of speech. In 1960, G. Fant published the classic Acoustic Theory of Speech Production which laid the foundations for many of the different methods of speech analysis that followed. As a direct result of the significant advances that occurred in understanding the acoustics of speech production, and with the aid of high-speed digital computers, the method of analysis-by-synthesis was given new impetus at M.I.T. (Bell, Fujisaki, Heinz, Stevens and House, 1961). A bank of 36 band-pass filters was used in their analysis. Another landmark was the pitch-synchronous analysis of voiced sounds as reported by Mathews, Miller and David (1961) at Bell Labs. They actually used analysis-by-synthesis on the spectrum of a single pitch period obtained by a Fourier analysis of the sampled waveform. In 1964, A.M. Noll introduced the cepstrum for the purpose of pitch extraction. The cepstrum was later used as the basis for a formant tracking system (Schafer and Rabiner, 1970). This very brief review gives a representative sample of the ideas and methodologies that have had a definite effect on the types of speech analysis that many speech researchers have chosen to pursue. A more complete review can be found in Flanagan (1972).

1.2 Linear Prediction

The past two years have witnessed a surge of interest on the part of the speech community in a method of analysis known alternately as predictive coding, linear prediction, Prony's method, inverse filtering formulation, etc. This surge of interest has

been also accompanied by an air of confusion. Two main reasons for this confusion are:

(1) A lack of exposition on the similarities and differences between different formulations.

(2) A resurfacing of some of the problems (e.g. windowing, preemphasis, etc.) associated with accepted methods for computation of short-time spectra.

We shall attempt, in this report, to deal with these problems by relating a few of these formulations to each other.

Let us first discuss what these formulations have in common. As far as we can ascertain, all the methods we have inspected have exactly one thing in common: they all assume that at a particular instant in time, a speech sample s(nT) can be approximated by a linearly weighted summation of the past p samples, where p is some integer.

$$s(nT) \cong \sum_{k=1}^{p} a_k \, s(nT-kT)$$

or

$$s_n \cong \sum_{k=1}^{p} a_k \, s_{n-k} \, , \qquad (1\text{-}1)$$

where T is the sampling interval, n is the sample number, and a_k, $1 \le k \le p$, are the weights. Equivalently, given p samples of a speech signal, the following sample can be predicted approximately by a linear summation of the p known samples. Hence the term "linear prediction". Henceforth we shall use the term "linear prediction" as a generic name for any method that makes an assumption equivalent to that in (1-1).

The problem at hand, as put forth by linear prediction, is to compute a set of predictor coefficients a_k such that (1-1) holds optimally over a specified period of time. It is in computing the set of coefficients a_k that different formulations of linear prediction have evolved.

The assumption in (1-1) could be made for any signal, be it speech or not. The reason that this assumption works well for speech is that it is based on a model of speech production which has been shown to work quite well in analysis-synthesis systems (Fant, 1960). Basically, the model assumes an all-pole transfer function of the combined effects of the glottal source, the

vocal tract and radiation. These poles can be computed by solving a polynomial in z with coefficients a_k. A more detailed description of this model is given in Chapter II.

Theoretically there exist an unlimited number of ways in which to compute the coefficients a_k. However, we shall initially limit our discussion to three formulations which we feel to be representative of the possible methods of analysis, and which raise some interesting issues. We shall describe briefly each of the formulations and give representative references on each without attempting to give a complete bibliography. The three methods will be given mnemonic names for ease of reference.

Exact Method

This method assumes that:

(a) The signal is defined for exactly 2p consecutive values.

(b) A speech sample can be predicted exactly from the past p samples, and that

(c) This holds for the trailing p consecutive samples.

These assumptions are represented by the following set of equations:

$$\sum_{k=1}^{p} a_k \, s_{n-k} = s_n, \quad n=0,1,\ldots,p-1. \tag{1-2}$$

These are p equations in p unknowns which in general can be solved for the coefficients a_k, $1 \le k \le p$.

Covariance Method

This method assumes that:

(a) The signal is defined for p+N consecutive values, where N is some integer.

(b) A speech sample can be approximately predicted from the past p samples, and that

(c) This holds for the trailing N consecutive samples.

(d) The total-squared error between the real signal and its predicted value is minimized over the N consecutive samples. (Some prefer to use the mean-squared error instead of total-squared error. The difference in this case is a division by a constant N which does not affect the results of minimization.)

The minimization of error results in the following set of equations (detailed derivation is shown in Section 3.1):

$$\sum_{k=1}^{p} a_k \phi_{ik} = \phi_{i0}, \quad i=1,2,\ldots,p \tag{1-3}$$

where

$$\phi_{ik} = \sum_{n=0}^{N-1} s_{n-i}\, s_{n-k} \,. \tag{1-4}$$

Again we have p equations in p unknowns which can be solved to obtain the coefficients a_k, $1 \le k \le p$. The coefficients ϕ_{ik} form a covariance matrix, hence the name "Covariance Method." Equations such as (1-3) are known in least-squares terminology as the normal equations of the process (Hildebrand, 1956, p. 260). In this case we shall call (1-3) the Covariance normal equations, or alternately the Covariance normal matrix equation.

Autocorrelation Method

The assumptions made in this method are:

(a) The signal is defined for all time such that it is identically zero outside a portion of the signal N samples long, where N is some integer. This is equivalent to multiplying the speech signal by a finite window of length N.

(b) Each sample can be approximately predicted from the past p samples, and that

(c) This is true for all time.

(d) The total-squared error between the actual signal and its predicted value is minimized for all time.

The minimization of error results in the following set of equations (the derivation is given in Section 3.1):

$$\sum_{k=1}^{p} a_k R_{|i-k|} = R_i \,, \quad i=1,2,\ldots,p \tag{1-5}$$

where

$$R_i = \sum_{n=0}^{N-1-|i|} s_n\, s_{n+|i|} \,. \tag{1-6}$$

Again (1-5) forms p equations with p unknowns to be solved for the coefficients a_k.

The R_i are autocorrelation coefficients of the signal. The coefficients $R_{|i-k|}$ form a special matrix which we shall call the autocorrelation matrix (as opposed to the covariance matrix in the Covariance method). Also, we shall call equations (1-5) the Autocorrelation normal equations or alternately the Autocorrela-

tion normal matrix equation.

As we shall see in Chapter IV, there are other possible formulations for the Covariance and Autocorrelation methods. The assumptions made above do not all apply in the other formulations. However, all Covariance-type formulations have (1-3) in common, and all Autocorrelation-type formulations have (1-5) in common, but (1-4) and (1-6) will not necessarily apply.

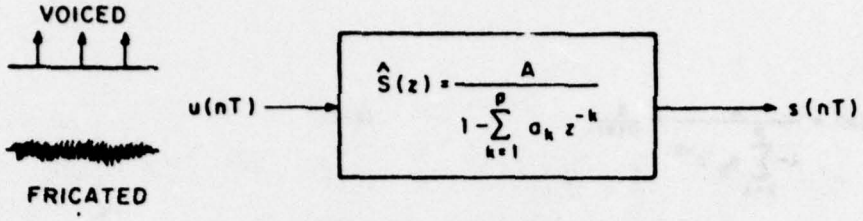
This concludes our brief description of each of three formulations for linear prediction. Now, we shall relate the work of some researchers to these three methods. The so-called Prony's method (Hildebrand, 1956, p. 378) or the exponential approximation method is equivalent to the Exact method for N = p and to the Covariance method for N≥p. A paper by Atal and Hanauer (1971),

For the purposes of modeling speech production, we approximate the continuously-varying vocal tract shape by a discretely-varying vocal tract shape, i.e. a vocal tract whose shape changes at discrete time intervals. Such a time interval shall be called a "frame". Within a frame, the vocal tract shape is considered to be fixed and can be modeled by a linear time-invariant filter. This model of speech production has been used effectively in speech synthesis systems. In linear prediction the linear filter is restricted to be all-pole.

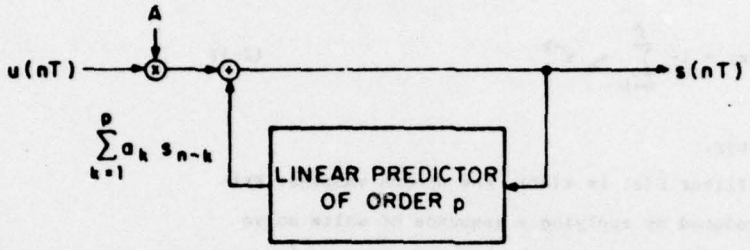
Thus, the model of speech production used in linear prediction consists of the following three assumptions:

(1) Within a short interval of time (on the order of 10-25 msec) the human vocal tract is assumed to be fixed in shape. We shall refer to such an interval as a "frame".

(2) Within any frame, we assume that the transfer function of the combined effects of the glottal flow, the vocal tract (including the oral and nasal cavities) and the radiation characteristic, can be modeled by a linear time-invariant all-pole filter with either a sequence of impulses or white noise (or a combination of both) as input (see Fig. 2-1).



(a) FREQUENCY-DOMAIN MODEL



(b) TIME-DOMAIN MODEL

Fig. 2-1. Discrete model of speech production as employed in linear prediction methods.

(3) The speech signal can be considered as the output of such an all-pole filter whose coefficients change at discrete intervals of time (on the order of 10 msec).

Below we shall focus our attention on a single frame where the all-pole filter is assumed to be time-invariant. Fig. 2-1a shows a schematic of the model in the frequency domain. The complex variable z is defined by:

$$z = e^{sT} = e^{(\sigma+j\omega)T}$$

where $s = \sigma+j\omega$ is the Laplace operator,

$\omega = 2\pi f$ is the radian frequency in rad/sec,

σ is the damping factor in rad/sec,

$T = \frac{1}{f_s}$ is the sampling interval in seconds,

and f_s is the sampling frequency in Hz.

• • •

Figure 2-1a is interpreted as follows: Speech is either voiced, fricated, or both. (Throughout this report we shall assume that aspiration is a kind of frication.) Voiced speech is produced by applying a sequence of impulses, spaced at the pitch period, to a digital filter of the form:

$$\hat{S}(z) = \frac{A}{1-\sum_{k=1}^{p} a_k z^{-k}} = \frac{A}{H(z)} \qquad (2\text{-}2)$$

where a_k, $1 \le k \le p$ are the filter coefficients,

A is a multiplicative gain factor that controls the signal amplitude,

and $$H(z) = 1 - \sum_{k=1}^{p} a_k z^{-k} \qquad (2\text{-}3)$$

is the inverse filter.

The output of the filter $\hat{S}(z)$ is $s(nT)$, the speech samples. Fricated speech is produced by applying a sequence of white noise samples, spaced T seconds apart, to a filter of the form $\hat{S}(z)$. Voiced fricatives are produced by a combination of voicing and frication. The filter $\hat{S}(z)$ represents the combined transfer function of the glottal flow, the vocal tract and radiation. The poles of the filter $\hat{S}(z)$ can be determined by solving for the roots of the polynomial in z in the denominator of $\hat{S}(z)$.

CHAPTER III

LINEAR PREDICTION ANALYSIS

In this chapter we shall derive in the time-domain the Covariance and Autocorrelation normal equations (1-3) and (1-5) and suggest algorithms for computing the predictor parameters. Given the normal equations, the minimum squared error is defined. The stability of the linear predictor, an important issue for speech synthesis, will then be examined for the three formulations of linear prediction. We then take a look at some autocorrelation-domain properties of linear prediction. A method for the computation of the gain factor A in $\hat{S}(z)$ will be specified.

3.1 Derivation of Covariance and Autocorrelation Normal Equations

Following the linear prediction speech production model described in Section 2.1 and represented by (2-6), we shall assume that a sampled speech signal s(nT) at time t=nT can be approximately predicted by a linear weighted summation of the past p samples. Let this approximation to s(nT) be $\tilde{s}(nT)$. We have:

$$\tilde{s}_n = \sum_{k=1}^{p} a_k s_{n-k} , \qquad (3-1)$$

where a_k, $1 \le k \le p$, is a set of real constants representing the predictor coefficients, and p is some integer whose value is determined as described in Sections 2.4 and 5.6.

Let the error between the actual value and the predicted value be given by e_n, where:

$$e_n = s_n - \tilde{s}_n$$

$$= s_n - \sum_{k=1}^{p} a_k s_{n-k} . \qquad (3-2)$$

The problem is to find a_k, $1 \le k \le p$, such that the error e_n is minimized in some sense over the desired range of signal samples. Both the Covariance and Autocorrelation methods employ a least-squares minimization procedure since it leads to a mathematically attractive solution. Denote the total-squared error by E, defined as:

$$E = \sum_n e_n^2 = \sum_n (s_n - \tilde{s}_n)^2 . \qquad (3-3)$$

The range over which the summation in (3-3) applies and the definition of s_n in that range is of importance. Indeed, this is exactly where the difference between the Covariance and Autocorrelation methods lies. However, let us first minimize E without specification of the range of the summation. Substituting (3-1) in (3-3) we obtain:

$$E = \sum_n (s_n - \sum_{k=1}^{p} a_k s_{n-k})^2 . \qquad (3-4)$$

The problem reduces to finding the condition that minimizes the total-squared error E with respect to a_k, $1 \le k \le p$. This condition is obtained by setting to zero the partial derivative of E with respect to each a_k:

$$\frac{\partial E}{\partial a_i} = \sum_n 2(s_n - \sum_{k=1}^{p} a_k s_{n-k})(-s_{n-i}) = 0, \qquad (3-5)$$

or,

$$\sum_n s_n s_{n-i} - \sum_n \sum_{k=1}^{p} a_k s_{n-k} s_{n-i} = 0, \quad 1 \le i \le p. \qquad (3-6)$$

Rearranging terms and interchanging summations we obtain:

$$\sum_{k=1}^{p} a_k \sum_n s_{n-k} s_{n-i} = \sum_n s_n s_{n-i}, \quad 1 \le i \le p. \qquad (3-7)$$

Equations (3-7) are known as the normal equations. For any definition of the signal s_n, (3-7) forms a set of p equations with p unknowns which can be solved for the predictor coefficients a_k. Now, we shall derive the Covariance and Autocorrelation normal equations from (3-7).

Covariance Normal Equations

Referring back to the assumptions of the Covariance method in Section 1.2, the summation over n in (3-3) and hence in (3-7) must go over N consecutive signal samples. Without loss of generality, we let the range of summation over n be: $n=0,1,\ldots,N-1$.

We can now write (3-7) as:

$$\sum_{k=1}^{p} a_k \phi_{ik} = \phi_{i0}, \quad i=1,2,\ldots,p \qquad (3-8)$$

where

$$\phi_{ik} = \sum_{n=0}^{N-1} s_{n-i} s_{n-k} . \qquad (3-9)$$

Note that (3-8) and (3-9) are identical to (1-3) and (1-4), and the derivation of the Covariance normal equations is complete. From (3-8) and (3-9) we note that values of s_n for $n=-p,\ldots,-1, 0,1,\ldots,N-1$, must be known. Therefore the signal s_n must be defined for p+N consecutive values, as stated in Section 1.2.

NOTES

This is the method used in the present study. (See Sec. III)

Autocorrelation Normal Equations

From the assumptions in Section 1.2 we can define the signal s_n as follows:

$$s_n = \begin{cases} \text{some sampled signal, } n=0,1,\ldots,N-1, \\ 0, \text{ otherwise.} \end{cases} \tag{3-10}$$

The windowed signal s_n is defined for all n: $-\infty<n<+\infty$. Equation (3-7) becomes:

$$\sum_{k=1}^{p} a_k \sum_{n=-\infty}^{\infty} s_{n-k}\, s_{n-i} = \sum_{n=-\infty}^{\infty} s_n\, s_{n-i}\,, \quad 1\le i\le p\,. \tag{3-11}$$

Substituting $m = n-i$ in (3-11) we obtain:

$$\sum_{k=1}^{p} a_k \sum_{m=-\infty}^{\infty} s_m\, s_{m+i-k} = \sum_{m=-\infty}^{\infty} s_m\, s_{m+i}\,, \quad 1\le i\le p. \tag{3-12}$$

By definition, the autocorrelation function R_i of the signal s_n is given by

$$R_i = \sum_{n=-\infty}^{\infty} s_n\, s_{n+|i|}. \tag{3-13}$$

and

$$R_{-i} = R_i\,. \tag{3-14}$$

Therefore, (3-12) reduces to:

$$\sum_{k=1}^{p} a_k\, R_{|i-k|} = R_i,\ i=1,2,\ldots,p. \tag{3-15}$$

Now, since s_n is defined in (3-10) to be identically zero for $n<0$ and $n\ge N$, (3-13) reduces to:

$$R_i = \sum_{n=0}^{N-1-|i|} s_n\, s_{n+|i|}\,. \tag{3-16}$$

Equations (3-15) and (3-16) are identical to (1-5) and (1-6), and the derivation of the Autocorrelation normal equations is complete.

3.2 Computation of Predictor Parameters

In each of the three formulations of linear prediction presented in Section 1.2 (eqs. 1-2, 3-8, 3-15), the predictor coefficients a_k, $1 \le k \le p$, can be computed by solving a set of p equations with p unknowns. There exist several standard methods for performing the necessary computations, e.g. the Gauss reduction or elimination method and the Crout reduction method (Hildebrand, 1956, pp. 428-434). These methods are general and can be used with the Exact, Covariance and Autocorrelation formulations. However, we note from the Covariance and Autocorrelation normal equations (3-8) and (3-15) that the matrix of coefficients in each case is a covariance matrix. The coefficients ϕ_{ik} in (3-8) form a typical covariance matrix and the coefficients $R_{|i-k|}$ in (3-15) form a special type of covariance matrix known as an autocorrelation matrix. A covariance matrix is symmetric and in general positive semidefinite, but in practice these covariance matrices are usually positive definite. Therefore, (3-8) and (3-15) can be solved more efficiently by the square-root method (Kunz, 1957, pp. 222-225). This method also requires about half the storage of the general methods. A similar method that does not employ the square root operation has been reported by Wilkinson and Reinsch (1971, pp. 9-30). Further reduction in storage and computation time is possible in solving the Autocorrelation normal equations because of their special form.

•

•

•

3.3 Minimum Total-Squared Error

The predictor coefficients a_k are determined such that the total-squared error E in (3-4) is minimized. After computation of the coefficients a_k using one of the methods mentioned in Section 3.2, one should be able to compute the minimum total-squared error E_p by substituting for the computed coefficients a_k in (3-4). (Note that there is no error criterion associated with the Exact method.) Thus:

$$E = \sum_n \left(s_n - \sum_{k=1}^{p} a_k s_{n-k} \right)^2$$

$$= \sum_n \left[s_n^2 - 2 s_n \sum_{k=1}^{p} a_k s_{n-k} + \sum_{k=1}^{p} \sum_{i=1}^{p} a_k a_i s_{n-k} s_{n-i} \right]$$

$$= \sum_n s_n^2 - 2 \sum_{k=1}^{p} a_k \sum_n s_n s_{n-k} + \sum_{k=1}^{p} a_k \sum_{i=1}^{p} a_i \sum_n s_{n-k} s_{n-i} .$$

Substituting (3-7), the condition for the minimization of E, and collecting terms, we obtain the minimum total-squared error E_p:

$$E_p = \sum_n s_n^2 - \sum_{k=1}^{p} a_k \sum_n s_n s_{n-k} . \qquad (3\text{-}18)$$

In particular, for the Covariance method, n ranges from 0 to N-1. Thus, substituting (3-9) in (3-18) we obtain the minimum total-squared error in the Covariance method:

$$E_p = \phi_{00} - \sum_{k=1}^{p} a_k \phi_{0k} . \quad \text{(Covariance Method)} \quad (3\text{-}19)$$

In the Autocorrelation method n ranges from -∞ to +∞. Substituting (3-13) in (3-18) we have:

$$E_p = R_0 - \sum_{k=1}^{p} a_k R_k . \quad \text{(Autocorrelation Method)} \quad (3\text{-}20)$$

We shall have the chance in Chapter V to discuss the behavior of this minimum error in the Autocorrelation method as a function of p and the autocorrelation function. In particular, we shall be interested in the normalized error V_p defined by:

$$V_p = \frac{E_p}{R_0} = \frac{\text{energy in the predictor error samples}}{\text{energy in the speech signal}} \qquad (3\text{-}21)$$

$$V_p = 1 - \sum_{k=1}^{p} a_k r_k , \qquad (3\text{-}22a)$$

here

$$r_k = \frac{R_k}{R_0} , \quad \text{for all } k , \qquad (3\text{-}22b)$$

and the samples r_k will be known as the <u>normalized autocorrelation function</u>. (Levinson (1947) uses the notation V, Markel (SCRL Mon., 1971) uses η, and Atal and Hanauer (1971) use ε for the normalized

error. We have chosen the letter V because of the possible usefulness of the normalized error in the indication of voicing.)
Note that dividing (3-15) by R_0 and using (3-22b) we obtain:

$$\sum_{k=1}^{p} a_k \, r_{|i-k|} = r_i \ , \quad 1 \le i \le p \ . \tag{3-23}$$

Equation (3-23) says that the predictor coefficients can also be computed using the normalized autocorrelation samples r_k. From (3-22b) and the fact that r_k is an autocorrelation function we have:

$$r_0 = 1$$

and $$|r_k| \le 1, \text{ for all } k. \tag{3-24}$$

The signal total energy R_0 can vary widely for different signals, which might cause round-off problems in trying to solve (3-15) in a digital computer with only integer arithmetic capability.
For such cases it would be useful to normalize the autocorrelation coefficients first by using (3-22b), and then solve for the a_k's using (3-23).

•

•

•

CHAPTER IV

SPECTRAL ESTIMATION AND ANALYSIS-BY-SYNTHESIS

In Chapter III the Covariance and Autocorrelation methods of linear prediction were derived from a time-domain formulation. In this chapter we shall show that the same normal equations can be derived from a frequency-domain formulation. It will become clear that linear prediction can be considered equally validly as either a time-domain or a frequency-domain type of analysis.

First, the Autocorrelation method is reinterpreted in terms of an inverse filter formulation. This leads directly to linear prediction analysis in the frequency domain. The Autocorrelation method is rederived from the spectral domain by approximating the signal short-time spectrum $P(\omega)$ by an all-pole power spectrum $\hat{P}(\omega)$. An error criterion between the two spectra is defined and minimized. The results are interpreted in terms of traditional methods of spectral analysis-by-synthesis. The Autocorrelation method is then reformulated in terms of a direct and an indirect method by relating to the corresponding methods of estimation of power spectra. An analogous reformulation of the Covariance method is derived from a generalized method of analysis-by-synthesis where the signal is assumed to be nonstationary and the two-dimensional short-time power spectrum $Q(\omega,\omega')$ is to be approximated by an all-pole two-dimensional spectrum $\hat{Q}(\omega,\omega')$.

• • •

4.1 Inverse Filter Formulation

The linear prediction error e_n was defined by (3-2), and is repeated here for convenience:

$$e_n = s_n - \sum_{k=1}^{p} a_k s_{n-k} . \qquad (3\text{-}2)$$

Since the signal s_n is defined for all time, then e_n is also defined for all time. Therefore, we can take the z-transform of (3-2) by multiplying both sides of the equation by z^{-n} and summing over all n • • •
The result is:

$$E(z) = S(z)\left[1 - \sum_{k=1}^{p} a_k z^{-k}\right]$$

$$= S(z)\, H(z), \qquad (1\text{-}1)$$

where $E(z)$ and $S(z)$ are the z-transforms of e_n and s_n, respectively, and $H(z) = 1 - \sum_{k=1}^{p} a_k z^{-k}$ was already defined in (2-3) as the inverse filter.

From (4-1), the error signal e_n can be interpreted as the output of a filter $H(z)$ whose input is s_n, as shown in Fig. 4-1.

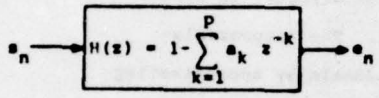


Fig. 4-1. The error sequence e_n as the output of an inverse filter $H(z)$.

Therefore, another way to view the error minimization problem in Section 3.1 is to solve for the parameters a_k of the inverse filter $H(z)$ which will minimize the energy $\sum_n e_n^2$ in the output error signal, for a given value of p. This is what Markel calls the inverse filter formulation (Markel, 1972).

Equation (4-1) can be solved for $S(z)$ to obtain:

$$S(z) = \frac{E(z)}{H(z)} = \frac{E(z)}{1 - \sum_{k=1}^{p} a_k z^{-k}} \qquad (4\text{-}2)$$

(4-2) is an exact equation. According to the speech production model described in Section 2.1, if the signal s_n is the vocal tract response due to a single pitch pulse, then the transfer function $S(z)$ can be approximated by an all-pole filter $\hat{S}(z)$ given by (2-2) and shown below:

$$\hat{S}(z) = \frac{A}{H(z)} = \frac{A}{1 - \sum_{k=1}^{p} a_k z^{-k}} \qquad (2\text{-}2)$$

Comparing (2-2) and (4-2) we conclude that $E(z)$ is approximated by another function

$$\hat{E}(z) = A \,,$$

which corresponds to a time-domain approximation $\hat{e}_n$ given by:

$$\hat{e}_n = A\, \delta_{n0} \,, \qquad (4\text{-}3)$$

where δ_{nm} is the Kronecker delta defined by (3-26).
$\hat{e}_n$ is just an impulse of magnitude A. Now, in order to conserve energy between $\hat{e}_n$ and e_n we must have

$$\sum_{n=-\infty}^{\infty} \hat{e}_n^2 = \sum_{n=-\infty}^{\infty} e_n^2 . \qquad (4\text{-}4)$$

After the minimization of the total-squared error, the right-hand side of (4-4) is equal to the minimum total-squared error E_p given by (3-20). The left-hand side of (4-4) is determined easily from (4-3), and we have:

$$A^2 = E_p = R_0 - \sum_{k=1}^{p} a_k R_k .$$

The result is identical to (3-37) which was derived by energy conservation between the signal s_n and the impulse response of $\hat{S}(z)$.

The above analysis assumed that the vocal tract was excited by a single pulse. The same results would be obtained if one assumed a white noise source excitation.

4.2 Error Minimization in the Spectral Domain

In this section we shall show that the Autocorrelation normal equations (3-15) can also be derived completely in the frequency domain. Before we proceed, we shall define the power spectrum of a transfer function Y(z) as the magnitude squared of Y(z) evaluated on the unit circle, i.e. $z = e^{j\omega T}$. Y(z) evaluated at $z = j\omega T$ will be denoted by Y(ω), so that the power spectrum is given by:

$$\text{Power Spectrum} = Y(\omega)\,\bar{Y}(\omega) = |Y(\omega)|^2 , \qquad (4\text{-}5)$$

where the over-bar denotes complex conjugate.
Let the power spectrum of $\hat{S}(z)$ be denoted by $\hat{P}(\omega)$, and of S(z) by P(ω), then:

$$\hat{P}(\omega) = |\hat{S}(\omega)|^2 = \frac{A^2}{\left|1 - \sum_{k=1}^{p} a_k e^{-jk\omega T}\right|^2} , \qquad (4\text{-}6a)$$

and

$$P(\omega) = |S(\omega)|^2 . \qquad (4\text{-}6b)$$

We shall call $\hat{P}(\omega)$ the linear prediction or approximate spectrum

and $P(\omega)$ the actual or signal spectrum. Methods for computing $P(\omega)$ and $\hat{P}(\omega)$ are given in Appendix C.

NOTES

Making use of Parseval's theorem (see Appendix A), the total-squared error E can be represented by:

$$E = \sum_{n=-\infty}^{\infty} e_n^2 = \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} |E(\omega)|^2 \, d\omega$$

$$= \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} P_e(\omega) \, d\omega \quad , \tag{4-7}$$

where $P_e(\omega)$ is the error power spectrum.

From linear system theory, we have from Fig. 4-1:

$$P_e(\omega) = P(\omega) \, |H(\omega)|^2 \quad , \tag{4-8}$$

where $H(\omega)$ is equal to $H(z)$ evaluated for $z = e^{j\omega T}$.

Substituting (4-8) in (4-7) we have:

$$E = \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} P(\omega) \, H(\omega) \, \bar{P}(\omega) \, d\omega, \tag{4-9}$$

$$= \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} P(\omega) \left[1 - \sum_{k=1}^{p} a_k \, e^{-jk\omega T}\right]\left[1 - \sum_{k=1}^{p} a_k \, e^{jk\omega T}\right] d\omega.$$

Following the same procedure in Section 3.1, E is minimized by setting $\frac{\partial E}{\partial a_i} = 0$, $1 \le i \le p$:

$$\frac{\partial E}{\partial a_i} = \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} P(\omega) \left[-e^{-ji\omega T}\left(1 - \sum_{k=1}^{p} a_k e^{jk\omega T}\right) - e^{ji\omega T}\left(1 - \sum_{k=1}^{p} a_k e^{-jk\omega T}\right)\right] d\omega = 0$$

or $$\frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} P(\omega) \left[\cos(i\omega T) - \sum_{k=1}^{p} a_k \cos((i-k)\omega T)\right] d\omega = 0.$$

There are different representations of the error E which are again used in section VI, with modfications to define measures, of distortion.

Interchanging integration and summation we have:

$$\sum_{k=1}^{p} a_k \left[\frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} P(\omega) \cos((i-k)\omega T) d\omega\right] = \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} P(\omega) \cos(i\omega T) \, d\omega, \quad 1 \le i \le p. \tag{4-10}$$

We know that the autocorrelation function $R(kT)$ is defined as the inverse Fourier transform of the power spectrum, i.e.

$$R_k = \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} P(\omega)\, e^{jk\omega T}\, d\omega, \qquad (4\text{-}11a)$$

or

$$R_k = \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} P(\omega)\, \cos(k\omega T)\, d\omega. \qquad (4\text{-}11b)$$

(4-11b) follows from (4-11a) because the power spectrum is a real and even function of frequency. Substituting (4-11b) in (4-10) and noting that $R_{-k} = R_k$, we have:

$$\sum_{k=1}^{p} a_k\, R_{|i-k|} = R_i\,, \quad 1 \le i \le p\,, \qquad (4\text{-}12)$$

which are the same Autocorrelation normal equations as (3-15).

The minimum total-squared error E_p can be obtained by using (4-10) and (4-11) in (4-9). The answer can be shown to be equal to

$$E_p = A^2 = R_0 - \sum_{k=1}^{p} a_k\, R_k\,, \qquad (4\text{-}13$$

which is identical to that given in (3-20) and (3-37).

The above derivation shows that, in the Autocorrelation method, the predictor parameters a_k can be determined if only the signal power spectrum is known. In fact all that is needed are the first p coefficients of the autocorrelation function, which can be computed either from the time signal (Section 3.1) or from the power spectrum as was shown above. The latter statement will be the basis for other formulations of the Autocorrelation method which are based on the idea of estimating the first p values of the autocorrelation function (see Section 4.4).

4.3 The Spectral Envelope and Analysis-by-Synthesis

We shall now interpret the minimization of error in the Autocorrelation method in terms of the estimation of the spectral envelope and in terms of analysis-by-synthesis.

From (2-2), H(z) can be written as:

$$H(z) = \frac{A}{\hat{S}(z)}\,,$$

and

$$H(\omega) = \frac{A}{\hat{S}(\omega)}\,. \qquad (4\text{-}14$$

NOTES

Substituting (4-14) in (4-9) we obtain:

$$E = \frac{A^2 T}{2\pi} \int_{-\pi/T}^{\pi/T} \frac{P(\omega)}{|\hat{S}(\omega)|^2} \, d\omega \, . \qquad (4\text{-}15)$$

$|\hat{S}(\omega)|^2$ is the approximate power spectrum $\hat{P}(\omega)$ as defined in (4-6a), and (4-15) reduces to:

$$E = \frac{A^2 T}{2\pi} \int_{-\pi/T}^{\pi/T} \frac{P(\omega)}{\hat{P}(\omega)} \, d\omega. \qquad (4\text{-}16)$$

Therefore, minimizing the total-squared error E is equivalent to the minimization of the integrated ratio of the signal power spectrum $P(\omega)$ to its approximation $\hat{P}(\omega)$. Another way to look at this is that if one is interested in approximating a power spectrum $P(\omega)$ by an all-pole spectrum $\hat{P}(\omega)$ then (4-16) is an error measure that can be used in optimizing the approximation.

•

•

•

6.21 Adequacy of the All-Pole Model

This issue has already been discussed in Section 2.3. We have argued there that the all-pole model seems to be quite adequate for speech synthesis. The question here is the adequacy of the model for formant extraction. For the purposes of speech recognition, for example, one would ideally want to be able to compute the transfer function of the vocal tract. This means that the antiformants as well as the formants may be needed. It is reasonable to assume that the all-pole model would be adequate for formant extraction of vowels. (This assumption is based on another assumption, namely that the glottal spectrum and radiation can be approximated by poles only.) However, for sounds such as nasals and fricatives, whose spectra are known to have antiformants, the all-pole model might not yield accurate results for the resonances of the vocal tract. Figure 6-3 shows the signal spectrum and the linear prediction spectrum (p=14) for the second [n] in the word "anyone" for a male speaker. The problem in looking at a spectrum

This represents the only potential insurmountable limitation to the use of the LPC technique for determining the performance of a voice channel.

like this is in deciding where the formants and antiformants are. There is no good way of making this decision in general, unless one has some knowledge about the system that produced the signal whose spectrum is under analysis. In fact, the spectral fit in Fig. 6-3 is very adequate, and it is quite reasonable to assume that some all-pole system has those characteristics. However, from our knowledge of the acoustics of the human speech production system, we know that if the spectrum in Fig. 6-3 is that of the sound [n], it must have zeros as well as poles. But even if we knew this, how would the linear prediction all-pole approximation help us in determining the values of the formants and antiformants? Some of the poles will correspond approximately to nasal formants, which can be obtained as described earlier in this section, but we know of no simple manner in which the antiformants can be determined from the poles of the linear prediction spectrum. The problem is that the same poles must approximate the effects of both the formants and the antiformants.

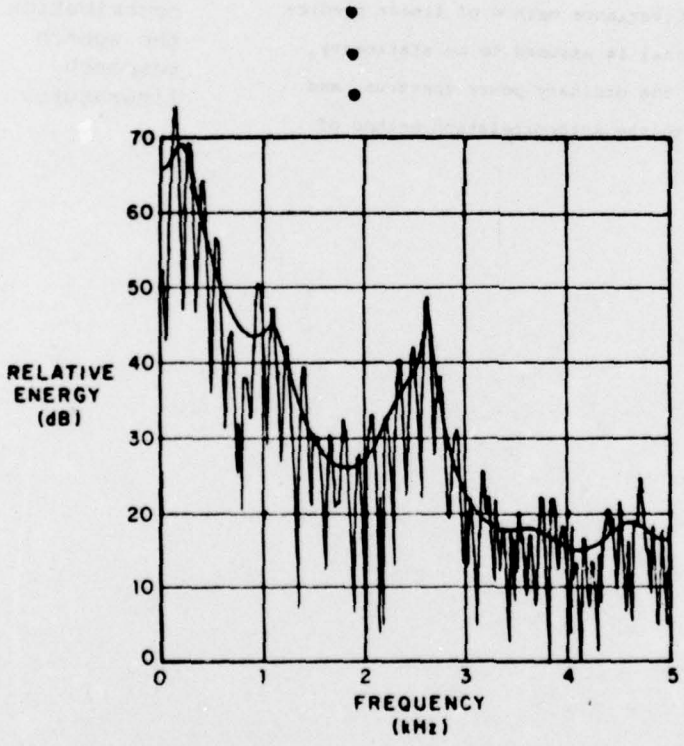


Fig. 6-3. Signal spectrum and linear prediction spectrum (p=14) for the second [n] in the word "anyone". Period of analysis is 25 msec.

CHAPTER VII

CONCLUSIONS

Linear prediction is an autocorrelation-domain analysis. Therefore, it can be approached either from the time or frequency domain. Although the actual computations are performed in the time domain, we chose to derive the most general formulations for linear prediction from the frequency domain because of the dominance of spectral analysis in speech research. We have shown that all least-squares methods of linear prediction can be derived from a single general concept, namely that of generalized analysis-by-synthesis. Here the 2D-spectrum (two dimensional spectrum) of a nonstationary signal (such as speech) is to be approximated by another 2D-spectrum, where the error to be minimized is proportional to the integral of the ratio of the signal spectrum to the approximate spectrum. This error criterion was shown to be very desirable for a good spectral envelope fit. In the special case when the approximate spectrum consists of poles only, the generalized method reduces to the general Covariance method of linear prediction. If, in addition, the signal is assumed to be stationary, the 2D-spectrum is replaced by the ordinary power spectrum, and the Covariance method reduces to the Autocorrelation method of linear prediction.

NOTES

Even though the technique is implemented in the time domain, frequency domain interpretations can also be obtained. However, it should be noted that it is more difficult, if not impossible, to obtain time domain interpretations from the usual spectral analyses that form a vast contribution to the speech research literature.